State-of-the-art in variable and functional form selection: update on splines

Aris Perperoglou [*] with Willi Sauerbrei and Georg Heinze for TG2 Stratos Initiative

[*]Visiting Professor Newcastle University
Data Science Director AstraZeneca

Context

Observational studies pose many design and statistical challenges

Valid observational research depends on careful study design, high data quality, appropriate statistical methods and accurate interpretation of results

The Problem

- Statistical methods have seen exponential advancements
  - diffusion of methodological innovation is slow
  - many developments are not applied in practice

- Even worse, “standard” analyses reported in the medical literature are often based on unrealistic assumptions or use inappropriate methods, casting doubt on their results and conclusions

- Analysts, reviewers, editors, readers and many more stakeholders and consumers need guidance for key issues in the design and analysis of observational studies

STRATOS Objectives

> Provide accessible and evidence-based guidance for key topics in the design and analysis of observational studies

> Guidance is intended for applied statisticians and other data analysts with varying levels of statistical education, experience and interests

Nine topic groups

Topic Group		Chairs
1	Missing Data	James Carpenter, Kate Lee
2	Selection of variables and functional forms in multivariable analysis	Georg Heinze, Aris Perperoglou, Willi Sauerbrei
3	Initial data analysis	Marianne Huebner, Saskia Le Cessie
4	Measurement error and misclassification	Laurence Freedman, Victor Kipnis
5	Study design	Suzanne Cadarette, Mitchell Gail
6	Evaluating diagnostic tests and prediction models	Ewout Steyerberg, Ben van Calster
7	Causal inference	Els Goetghebeur, Ingeborg Waernbaum
8	Survival analysis	Michal Abrahamowicz, Per Kragh Andersen, Terry Therneau
9	High-dimensional data	Lisa McShane, Joerg Rahnenfuehrer

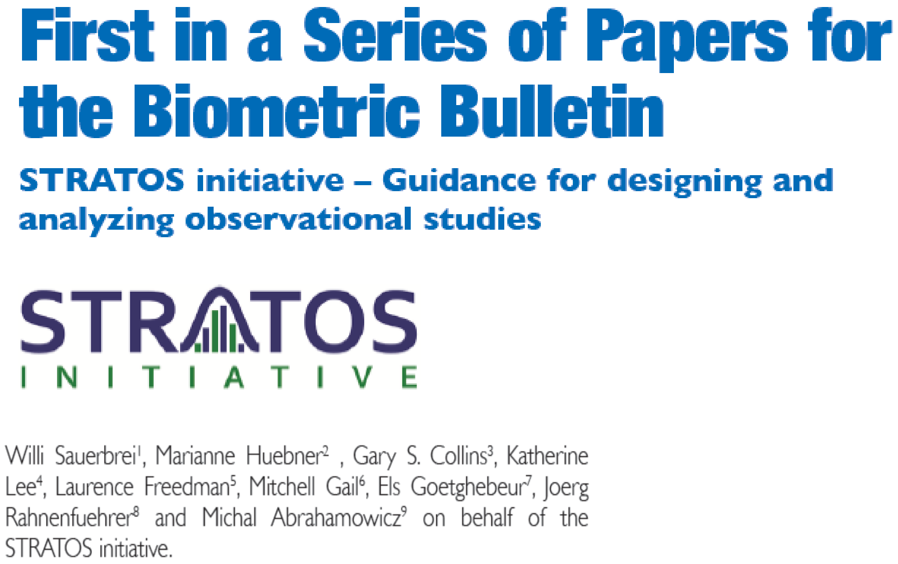
Eleven cross-cutting panels

Panel		Chairs and Co-Chairs	
MP	Membership	Chairs	James Carpenter, Willi Sauerbrei
PP	Publications	Chairs	Bianca De Stavola, Pamela Shaw
		Co-Chairs	Mitchell Gail, Petra Macaskill
GP	Glossary	Chairs	Simon Day, Marianne Huebner, Jim Slattery
WP	Website	Chairs	Joerg Rahnenfuehrer, Willi Sauerbrei
RP	Literature Review	Chairs	Gary Collins, Carl Moons
BP	Bibliography	Chairs	to be determined
SP	Simulation Studies	Chairs	Michal Abrahamowicz, Anne-Laure Boulesteix
DP	Data Sets	Chairs	Saskia Le Cessie, Maarten van Smeden
TP	Knowledge Translation	Chairs	Suzanne Cadarette
		Co-Chair	Catherine Quantin
CP	Contact Organizations	Chairs	Willi Sauerbrei
VP	Visualisation	Chairs	Mark Baillie

Guidance for analysis is needed for many stakeholders (analysts with different levels of knowledge, teachers, reviewers, journalists, ……)

Researchers

Consumers



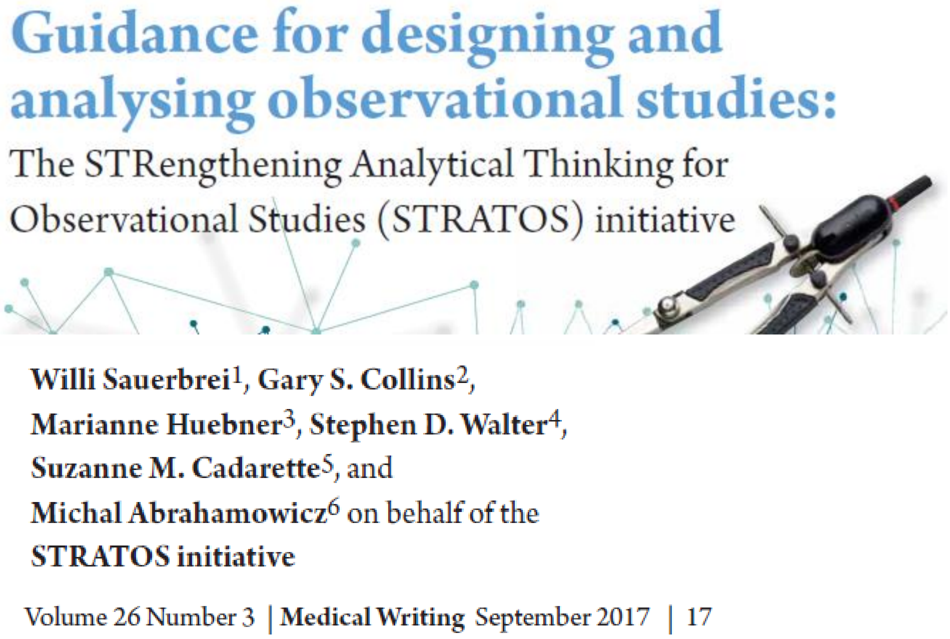
First in a Series of Papers for the Biometric Bulletin

STRATOS initiative – Guidance for designing and analyzing observational studies

STRATOS INITIATIVE

Willi Sauerbrei[1], Marianne Huebner[2] , Gary S. Collins[3], Katherine Lee[4], Laurence Freedman[5], Mitchell Gail[6], Els Goetghebeur[7], Joerg Rahnenfuehrer[8] and Michal Abrahamowicz[9] on behalf of the STRATOS initiative.

Short papers from all nine topic groups and the simulation panel have appeared



Guidance for designing and analysing observational studies:

The STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative

Willi Sauerbrei[1], **Gary S. Collins**[2], **Marianne Huebner**[3], **Stephen D. Walter**[4], **Suzanne M. Cadarette**[5], and **Michal Abrahamowicz**[6] on behalf of the **STRATOS initiative**

Volume 26 Number 3 | **Medical Writing** September 2017 | 17

Journal of the European Medical Writers Association (EMWA)

TG2
Selection of Variables and Functional Forms in Multivariable Analysis

Descriptive models: (TG2)
Capture the association of explanatory and outcome variables

Predictive modeling: (TG6)
Transparent (as opposed to black-box) prediction models, often with superior performance background knowledge can be easily inserted

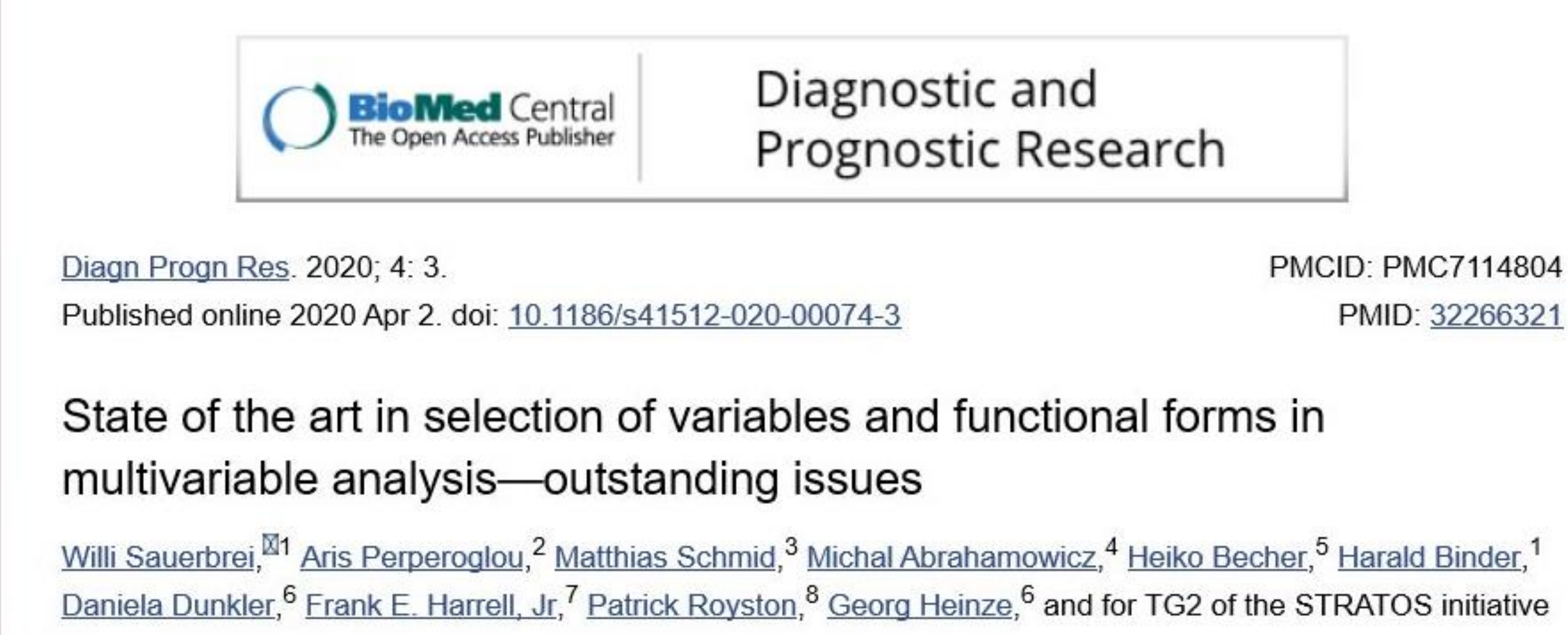
Explanatory modeling: (TG7)
Designed to estimate an identifiable causal effect of interest directly or for prediction of counterfactual outcomes

STRATOS INITIATIVE

Aims based on different levels of experience

Level-1:	> teach multivariable model building to non-statisticians > give recommendations
Level-2:	> summarize state-of-the-art and key issues > give recommendations
Level-3:	> evaluate what are the recommendable strategies and procedures for multivariable modelling building

State-of-the-art



BioMed Central
The Open Access Publisher

Diagnostic and Prognostic Research

Diagn Progn Res. 2020; 4: 3.
Published online 2020 Apr 2. doi: 10.1186/s41512-020-00074-3

PMCID: PMC7114804
PMID: 32266321

State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues

Willi Sauerbrei,✉[1] Aris Perperoglou,[2] Matthias Schmid,[3] Michal Abrahamowicz,[4] Heiko Becher,[5] Harald Binder,[1] Daniela Dunkler,[6] Frank E. Harrell, Jr,[7] Patrick Royston,[8] Georg Heinze,[6] and for TG2 of the STRATOS initiative

STRATOS
INITIATIVE

Further research needed:

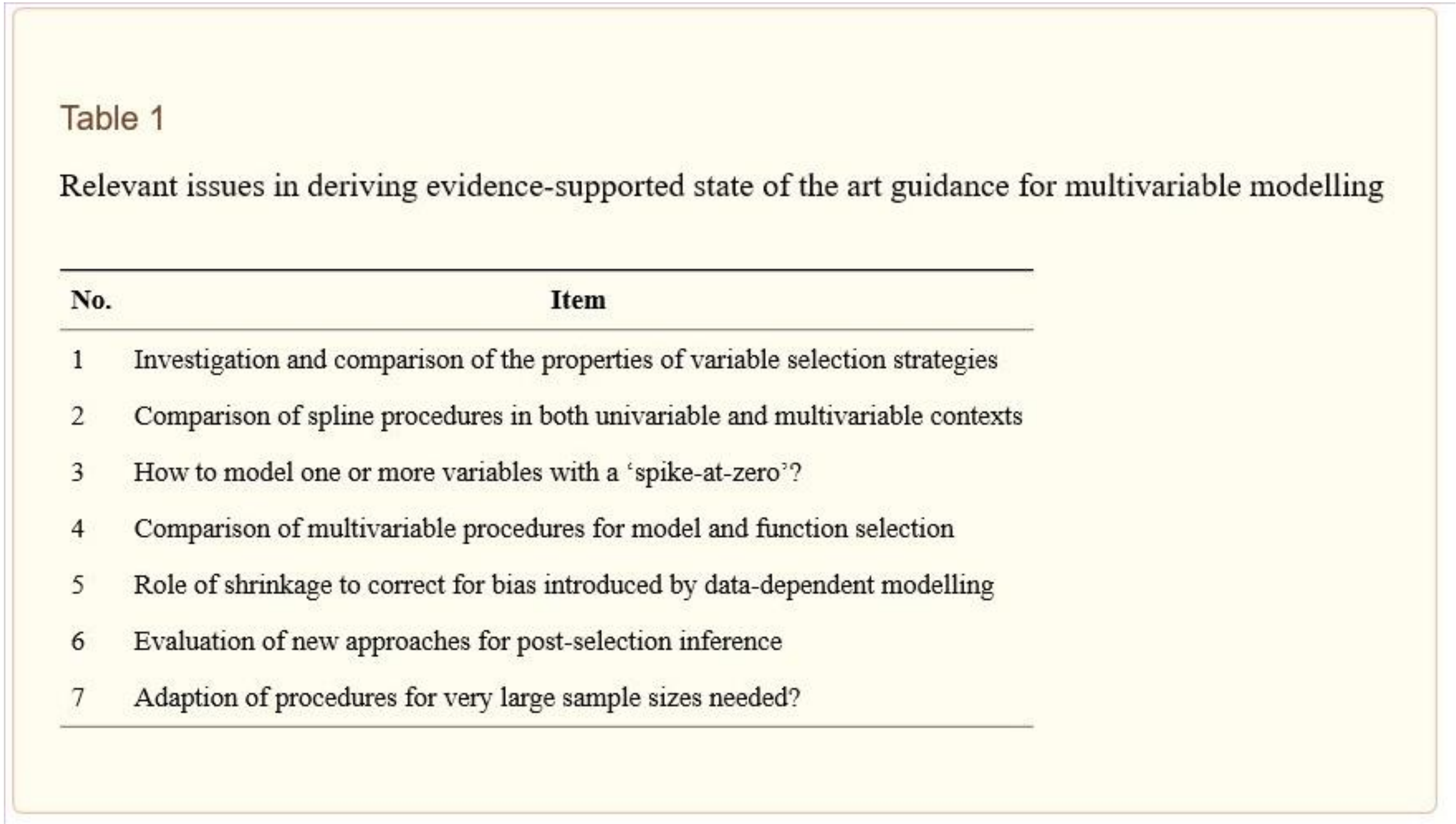
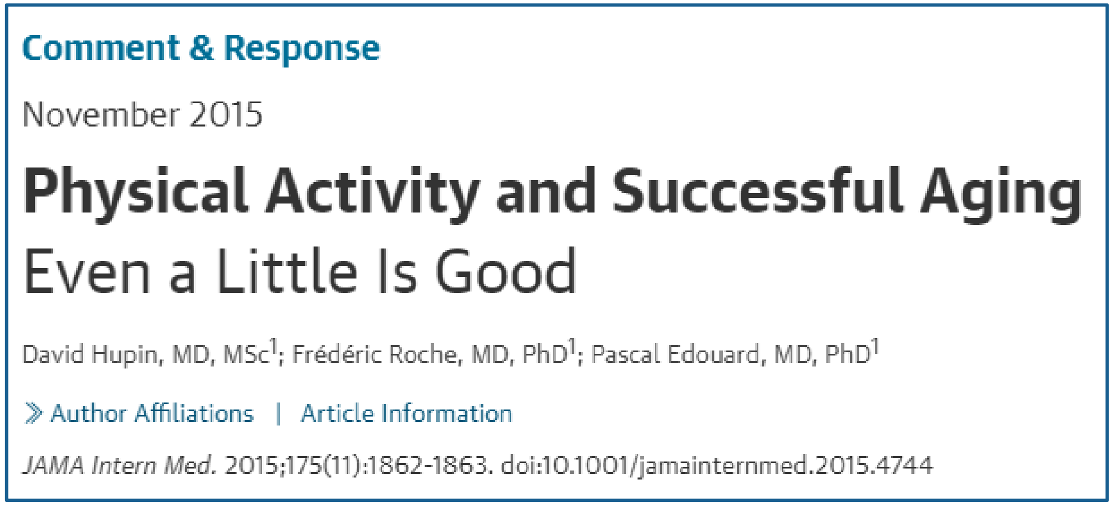


Table 1

Relevant issues in deriving evidence-supported state of the art guidance for multivariable modelling

No.	Item
1	Investigation and comparison of the properties of variable selection strategies
2	Comparison of spline procedures in both univariable and multivariable contexts
3	How to model one or more variables with a ‘spike-at-zero’?
4	Comparison of multivariable procedures for model and function selection
5	Role of shrinkage to correct for bias introduced by data-dependent modelling
6	Evaluation of new approaches for post-selection inference
7	Adaption of procedures for very large sample sizes needed?

Maybe we are overreacting:



Comment & Response

November 2015

Physical Activity and Successful Aging

Even a Little Is Good

David Hupin, MD, MSc[1]; Frédéric Roche, MD, PhD[1]; Pascal Edouard, MD, PhD[1]

» Author Affiliations | Article Information

JAMA Intern Med. 2015;175(11):1862-1863. doi:10.1001/jamainternmed.2015.4744

JAMA Internal Medicine (IF 15)

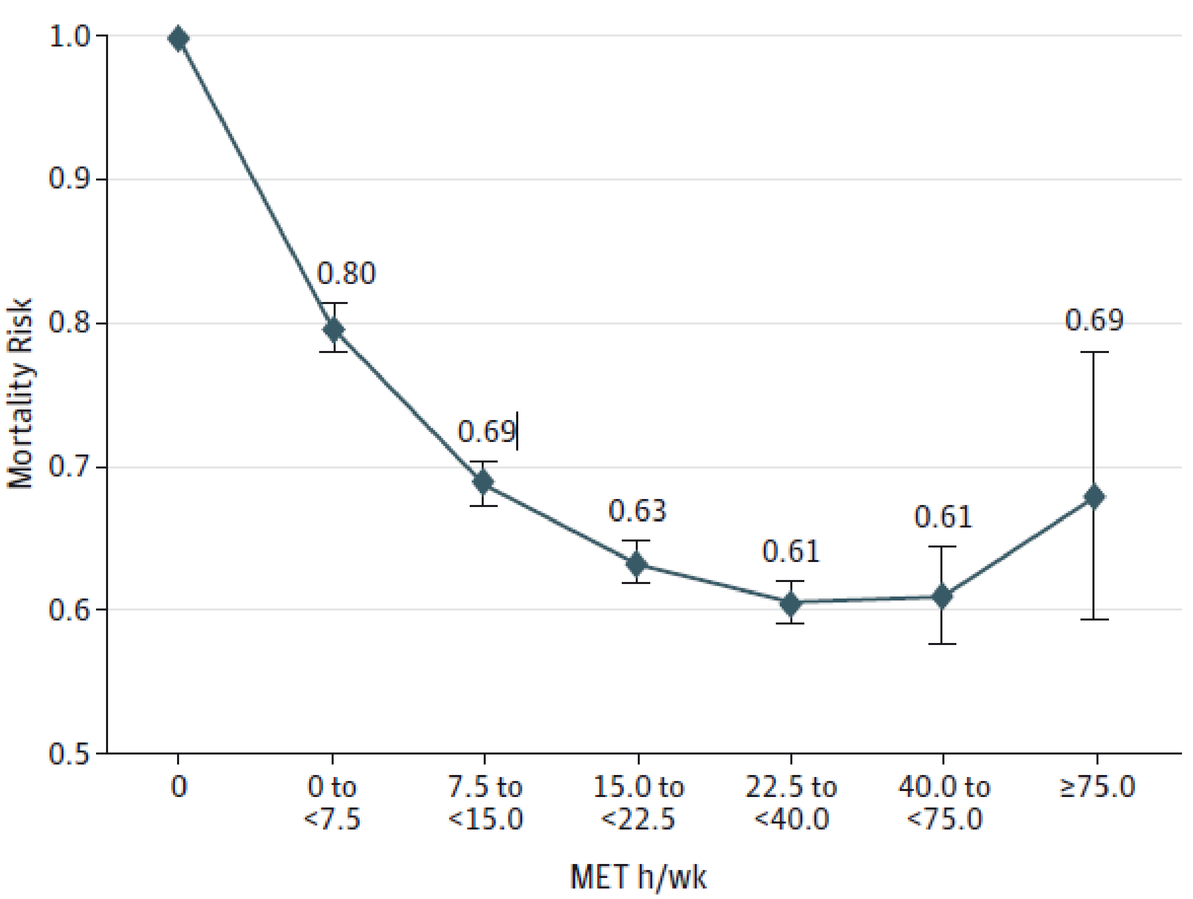
> N=666,137

> Main exposure: metabolic equivalent training (MET) in hours/week

> For the main analysis, MET was categorized into

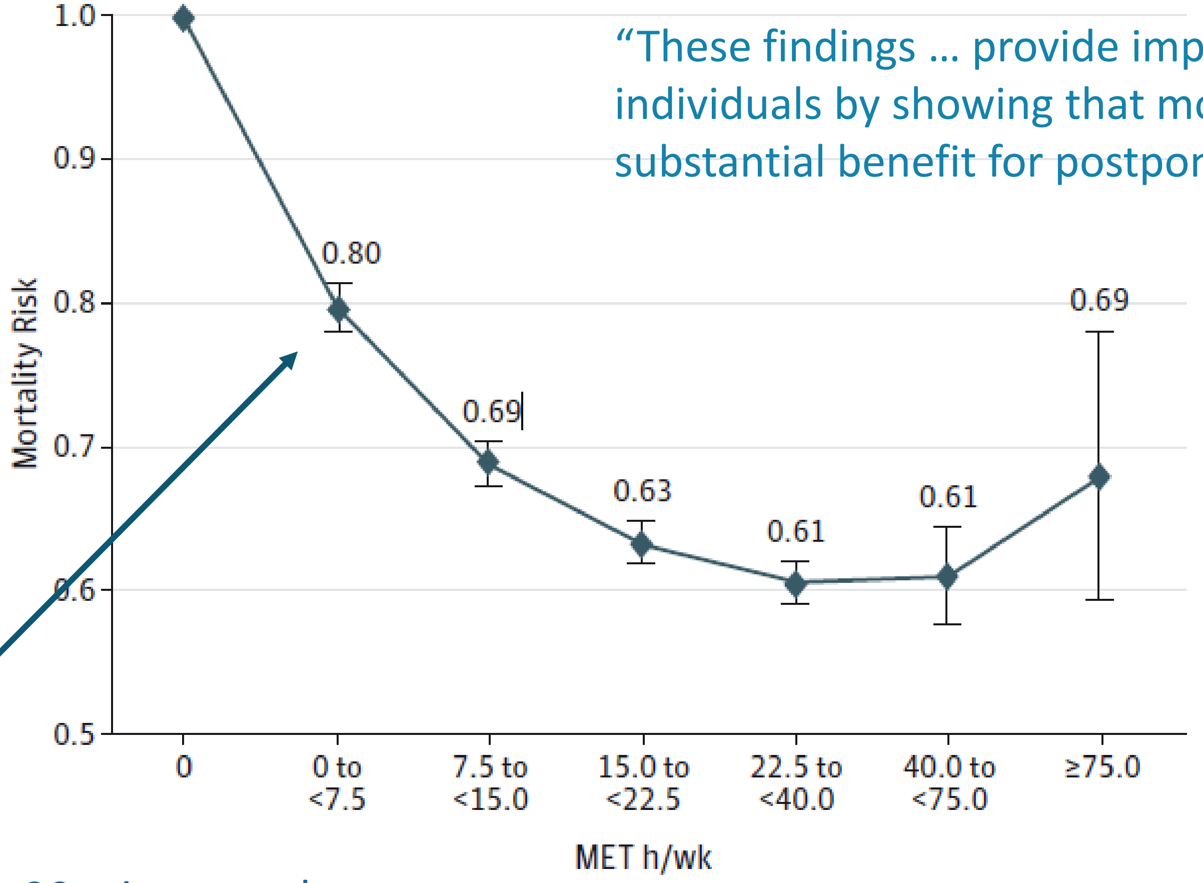
0 h/w, 0.2-7.5, 7.7-15, 15.2-22.5, 22.7-40, 40.2-75, 75.2+

Figure. Hazard Ratios (HRs) and 95% CIs for Leisure Time Moderate- to Vigorous-Intensity Physical Activity and Mortality



There is indeed a need for dichotomous decisions Treat / NotTreat, but that need does not justify dichotomisation/categorisation of covariates.

Figure. Hazard Ratios (HRs) and 95% CIs for Leisure Time Moderate- to Vigorous-Intensity Physical Activity and Mortality



"These findings … provide important evidence to inactive individuals by showing that modest amounts of activity provide substantial benefit for postponing mortality"

Effect of walking 16 seconds to 20 minutes a day

Level 1 guidance

Bone Marrow Transplantation

Editorial | Published: 01 October 2019

Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians

J. Gauthier, Q. V. Wu & T. A. Gooley

Bone Marrow Transplantation **55**, 675–680(2020) | Cite this article

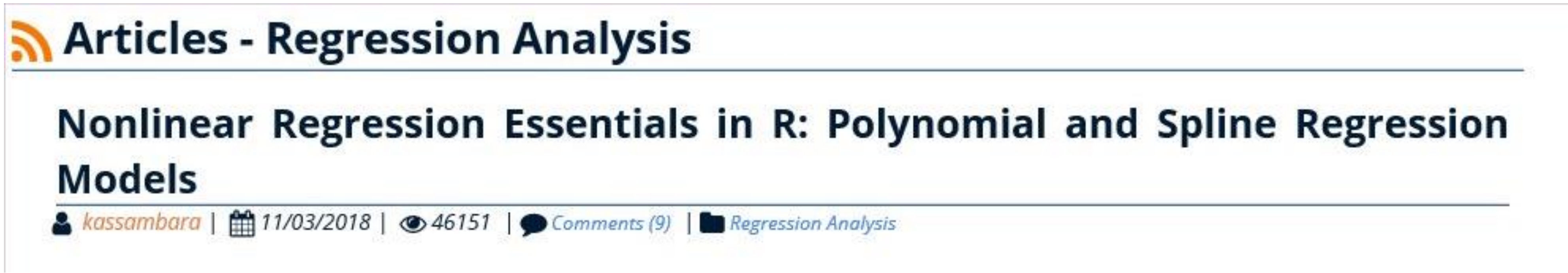
5528 Accesses | **3** Citations | **7** Altmetric | Metrics

Suggests using restricted spline instead of categorization

Very basic approach, no mention on how to choose number/place of knots

Only one mention of overfitting (when many knots are used)

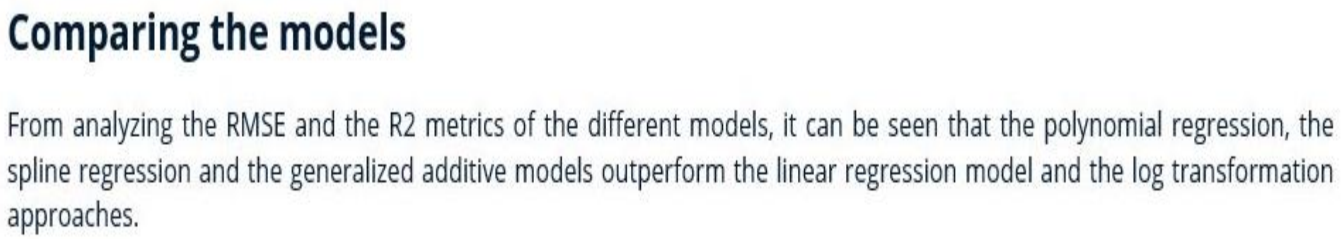
Level 0 guidance (online tutorial)



Articles - Regression Analysis

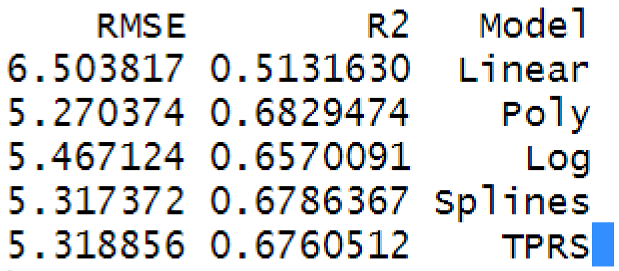
Nonlinear Regression Essentials in R: Polynomial and Spline Regression Models

kassambara | 11/03/2018 | 46151 | Comments (9) | Regression Analysis

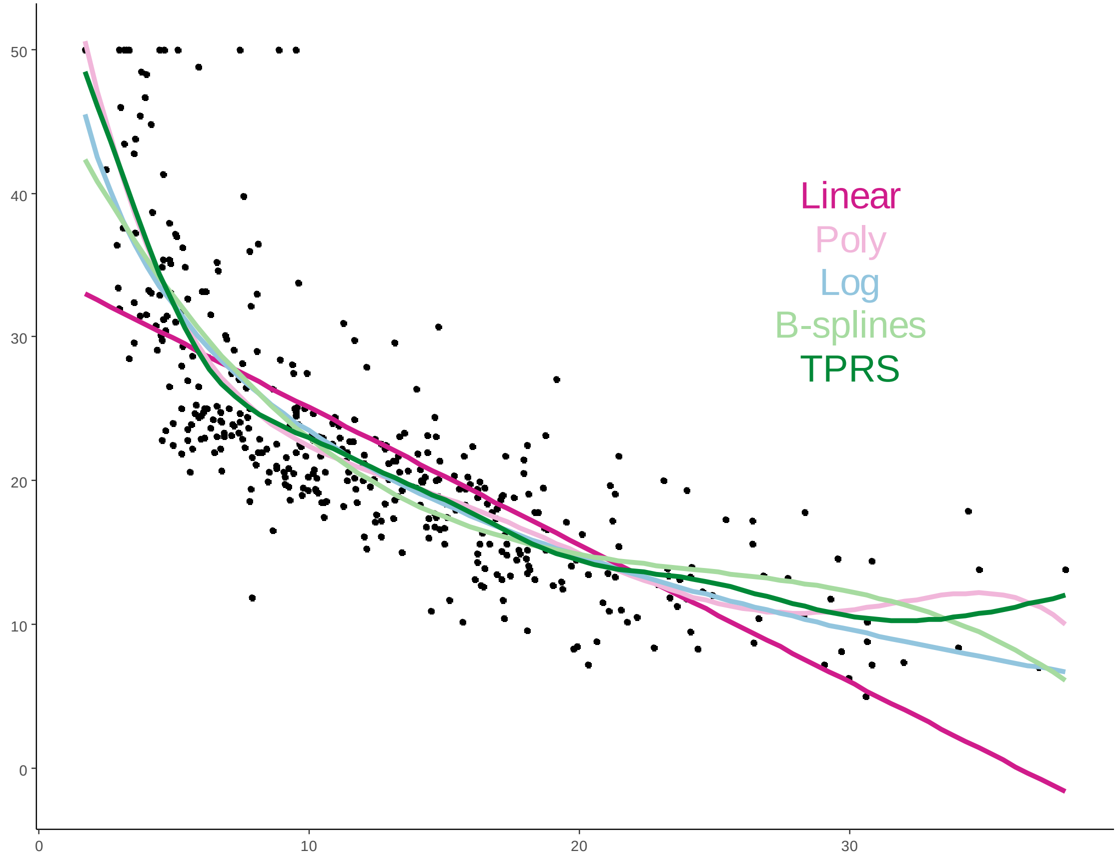


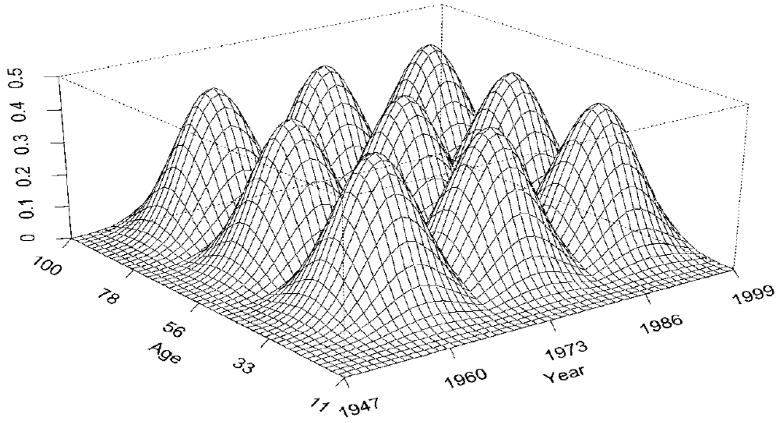
Comparing the models

From analyzing the RMSE and the R2 metrics of the different models, it can be seen that the polynomial regression, the spline regression and the generalized additive models outperform the linear regression model and the log transformation approaches.



```
     RMSE        R2   Model
6.503817 0.5131630  Linear
5.270374 0.6829474    Poly
5.467124 0.6570091     Log
5.317372 0.6786367 Splines
5.318856 0.6760512    TPRS
```



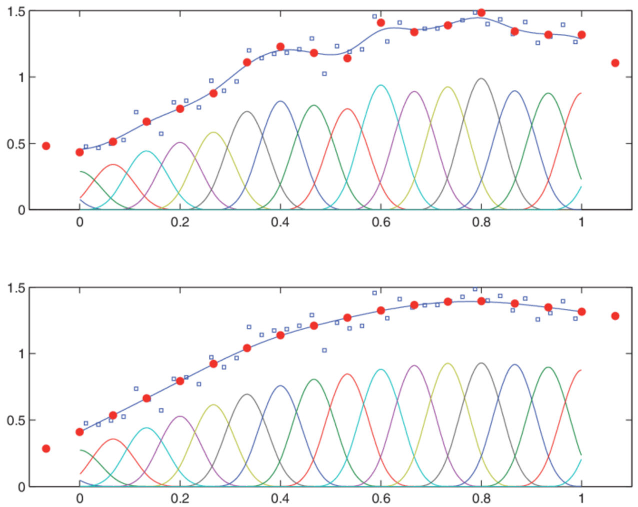


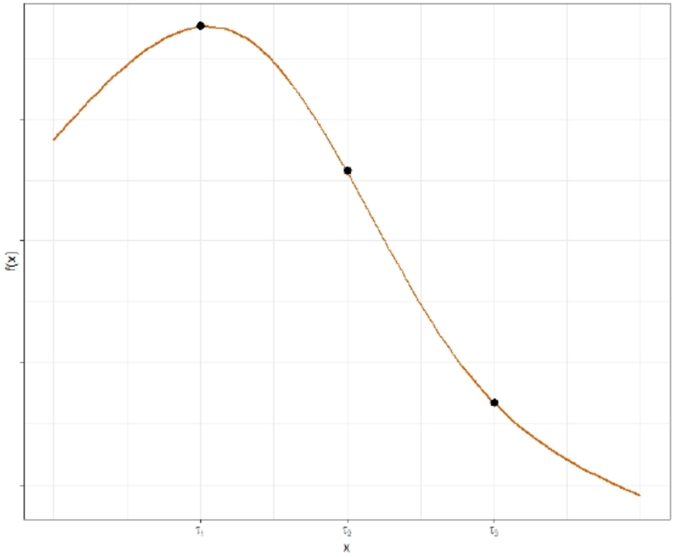
Splines are beautiful:

Set of piecewise polynomials, each of degree d

Joined together at a set of knots $\tau_1 \ldots \tau_k$

Continuous in value and sufficiently smooth at the knots

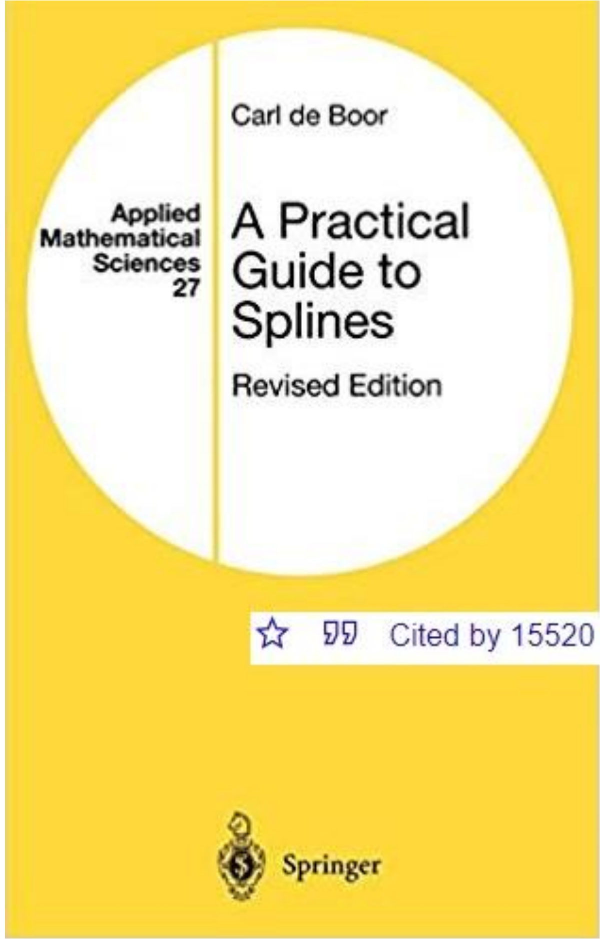




Spoiled for choice:

- Type of function (polynomial) and its degree → spline basis
  - Polynomial, cubic spline, natural, b-splines….
- Number and position of knots
- Regression splines or smoothing splines (penalised)
  - b-splines vs p-splines, thin plate regression splines, o-splines, m-splines
- Penalty weight, optimisation methods (AIC/BIC, GCV, REML), matrix of differences…

Some references:



Carl de Boor

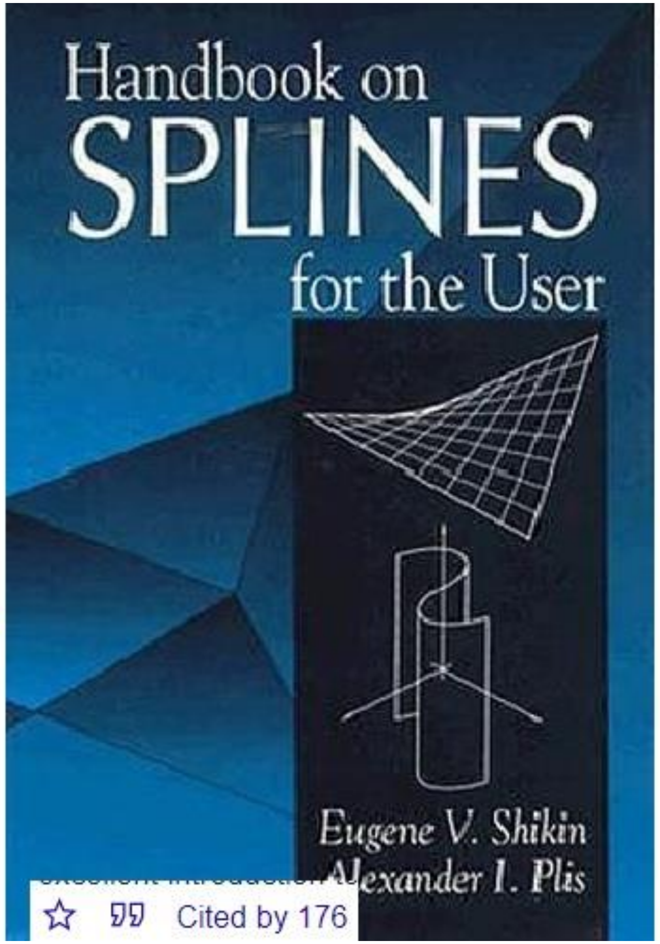
Applied Mathematical Sciences 27

A Practical Guide to Splines

Revised Edition

Springer

☆ 99 Cited by 15520



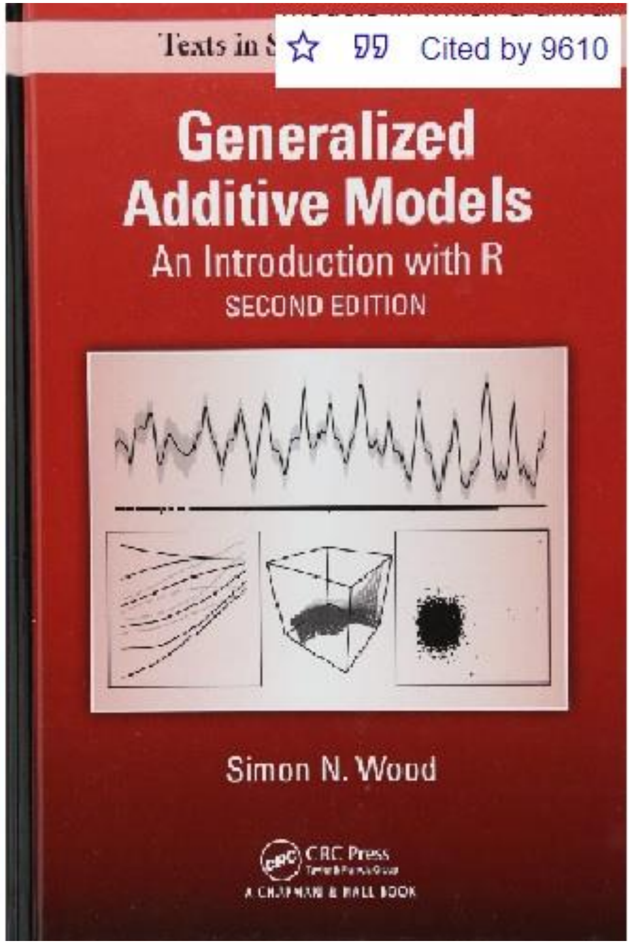
Handbook on SPLINES for the User

Eugene V. Shikin

Alexander I. Plis

excellent introduction to

☆ 99 Cited by 176



Texts in S

☆ 99 Cited by 9610

Generalized Additive Models

An Introduction with R

SECOND EDITION

Simon N. Wood

CRC Press

A CHAPMAN & HALL BOOK

The need for guidance:

> Splines can be daunting, especially due to the number of choices a researcher must make.

> Most researchers are not taught how to use splines.

> In many cases researchers use off the shelf software at default values of procedures.

> There is a lack of comparisons between different approaches.

Comparison of spline procedures

We would like to know:

> How results from various spline procedures differ from true function, and how does this depend on relevant parameters ?

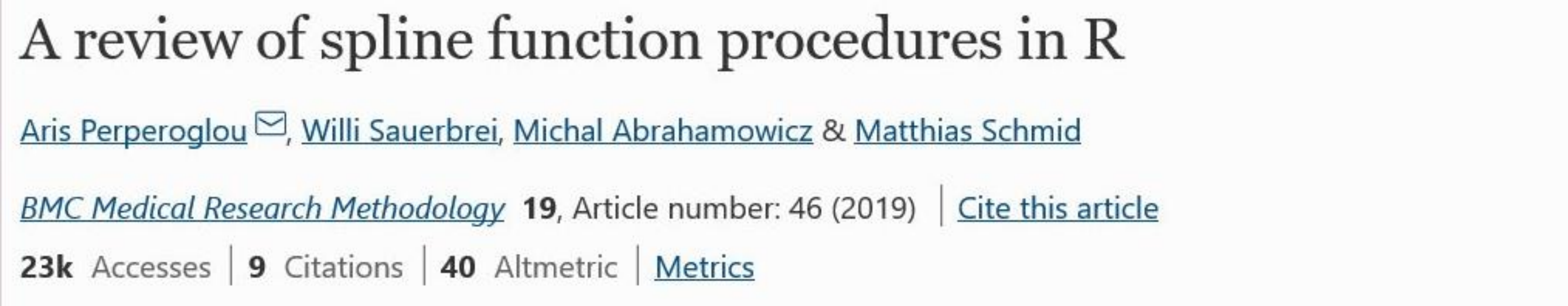
> Permitted complexity, usability for non-experts

> Multivariable context – multiple variables of mixed types

For level-1: **How to report results in a clinical paper?**

Just a supplementary figure, or main result?

Recommendations for typical contrasts to report?

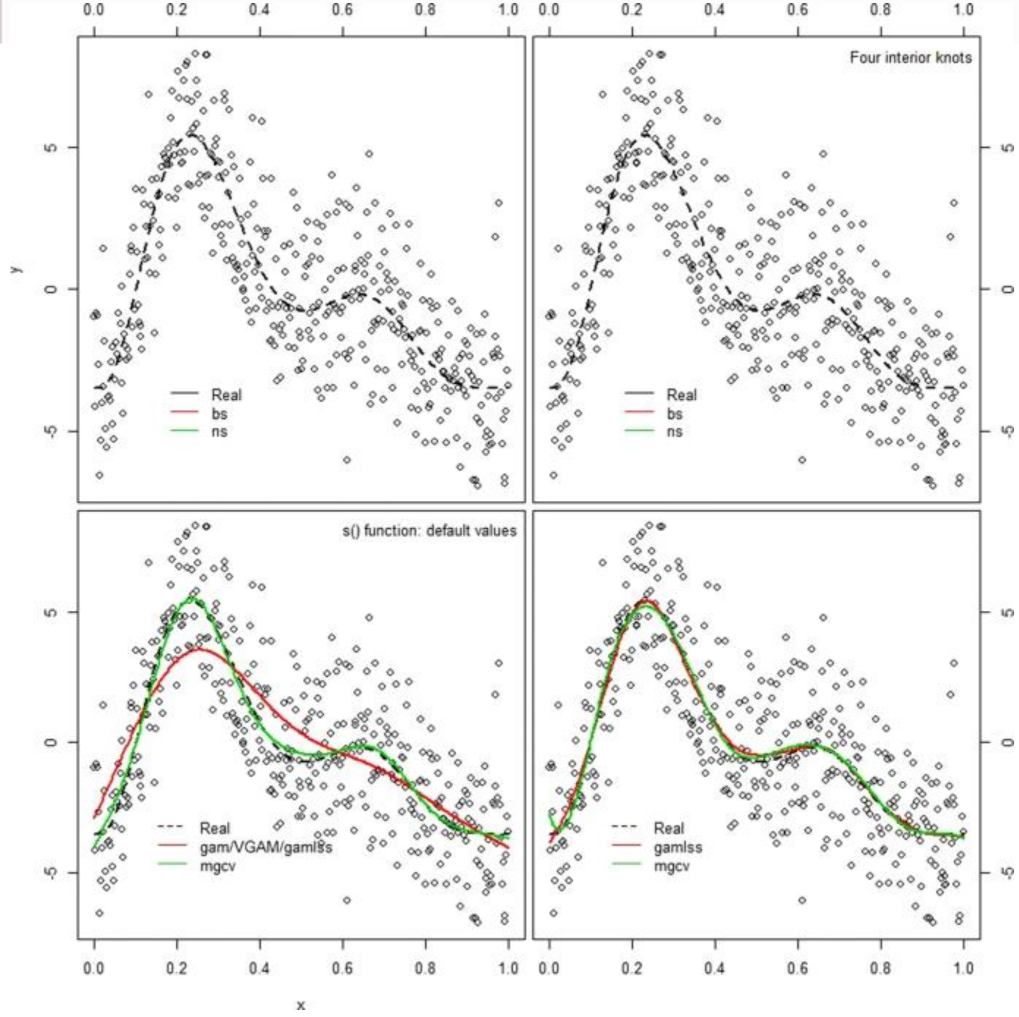


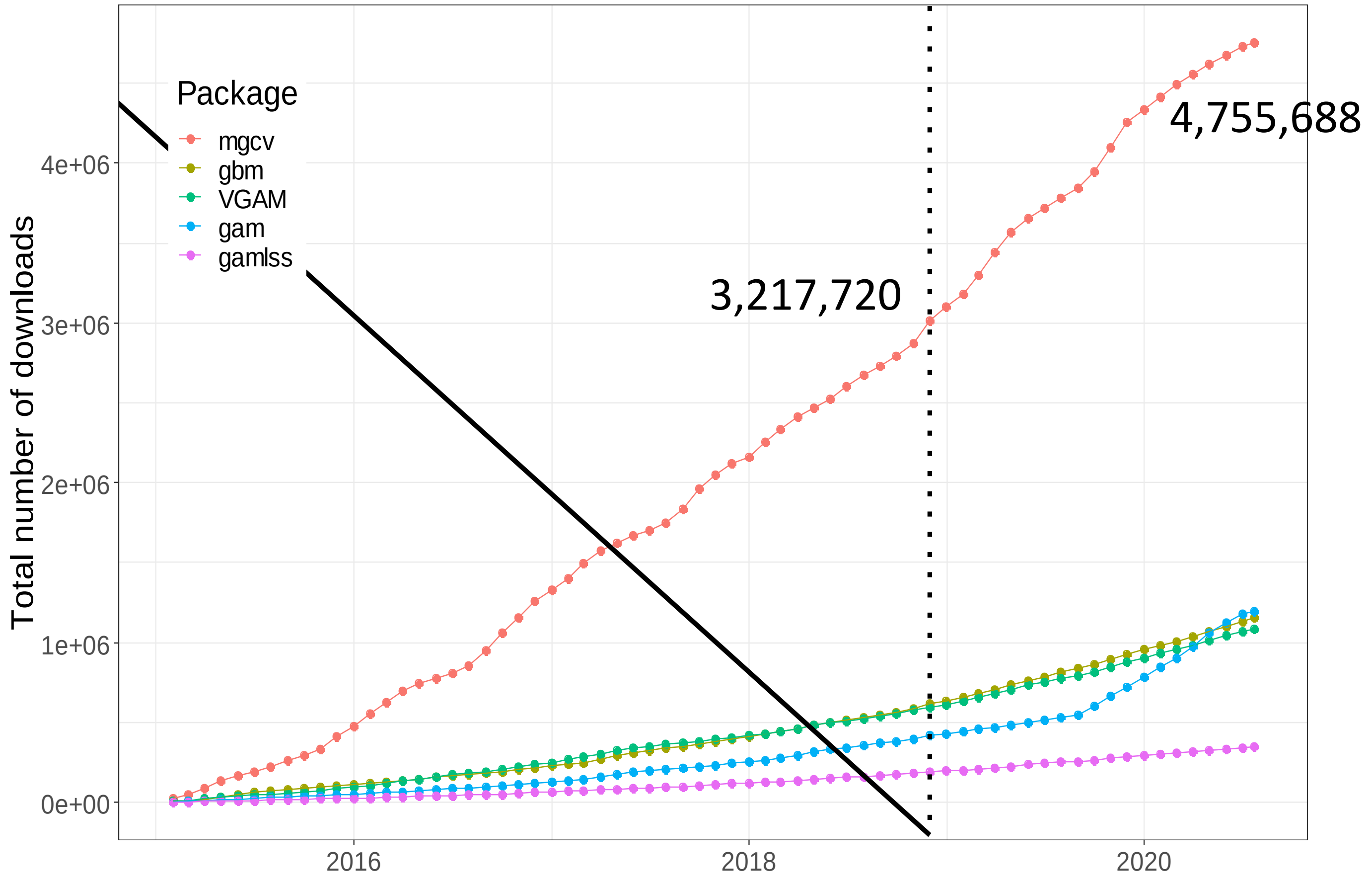
A review of spline function procedures in R

Aris Perperoglou ✉, Willi Sauerbrei, Michal Abrahamowicz & Matthias Schmid

BMC Medical Research Methodology **19**, Article number: 46 (2019) | Cite this article

23k Accesses | **9** Citations | **40** Altmetric | Metrics





Experts advice:

Frank Harrell Jr (RMS 2019) on restricted cubic splines:

> k Knots are specified in advance

> Choice of k depends on sample size

> For n>100 then k=5

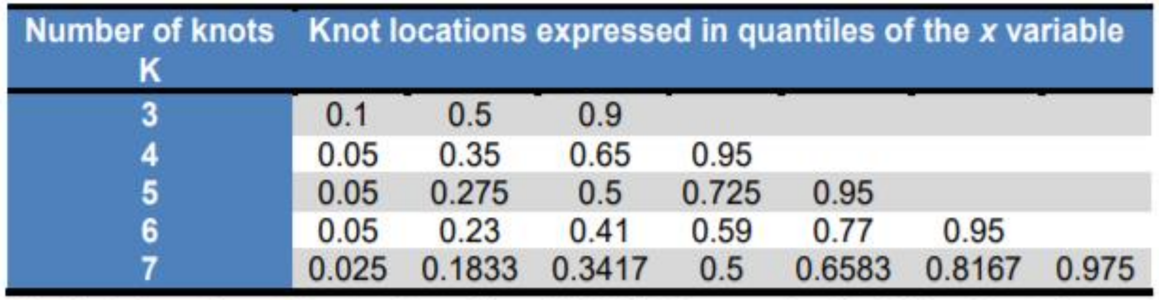
> For n<30 then k=3

> Often k=4 is enough

> Or use AUC t choose k

> Location is not crucial in most situations

as long as knots are where data exist – fixed quantiles



Number of knots K	Knot locations expressed in quantiles of the *x* variable						
3	0.1	0.5	0.9				
4	0.05	0.35	0.65	0.95			
5	0.05	0.275	0.5	0.725	0.95		
6	0.05	0.23	0.41	0.59	0.77	0.95	
7	0.025	0.1833	0.3417	0.5	0.6583	0.8167	0.975

Table 2. Location of knots. From Harrell (2001), Regression Modeling Strategies.

Eilers and Marx (Statistical Science 1996) on p-splines

> Regression on cubic b-splines

> Use large number of knots (10, 20, 50)

> Use a difference penalty (order 2 or 3) on the coefficients

> Tune smoothness with penalty weight (λ)

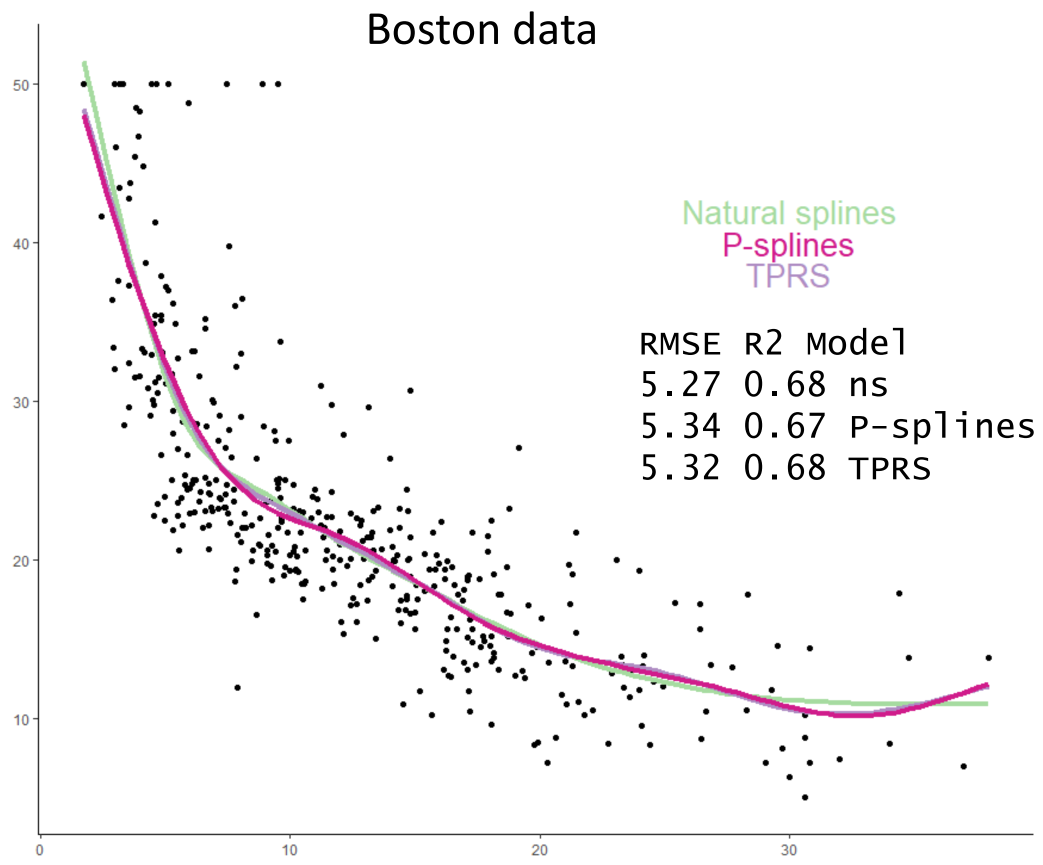
Simon Wood (A toolbox of smooths 2009) on thin plate regression splines

> Eigen based approach vs knots based

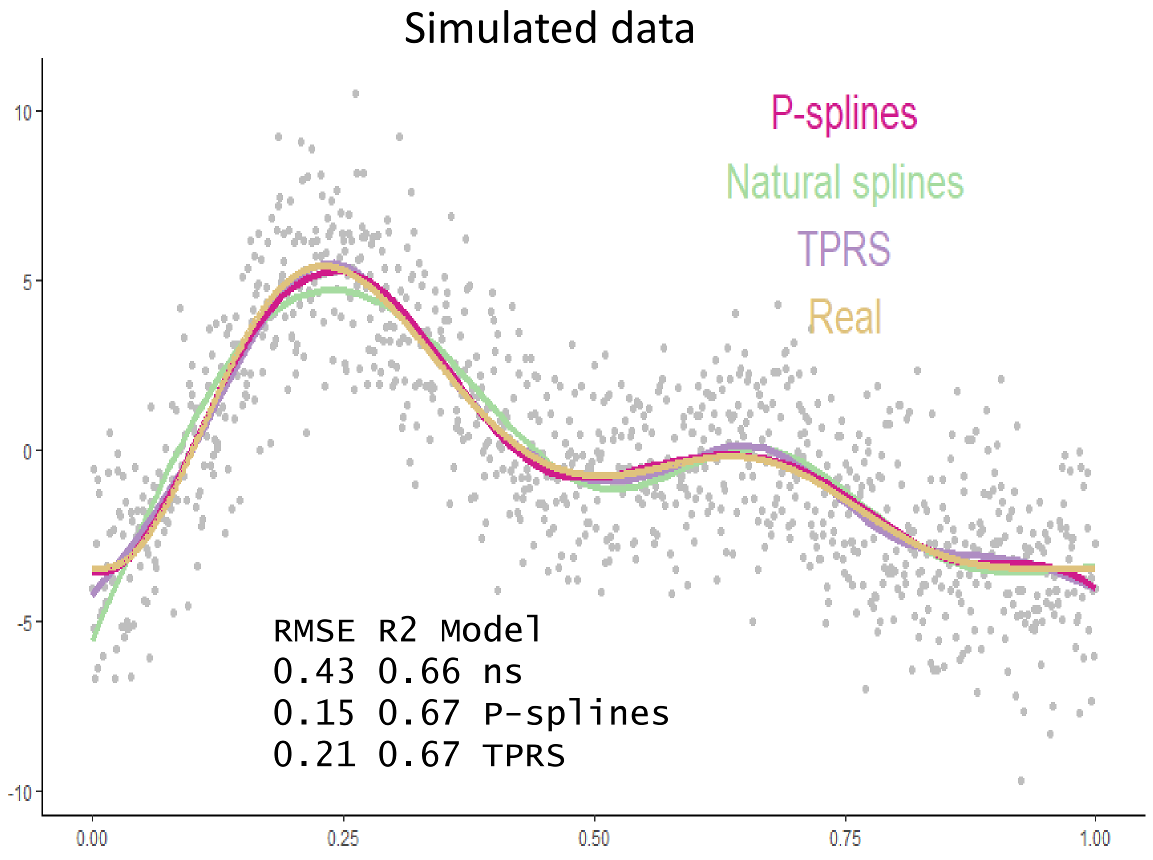
> Choose how many basis functions are to be used and then solve the problem of finding the set of this many basis functions that will optimally approximate a full spline.

> Default on mgcv 23 basis functions, GCV for optimisation

An example


Boston data

```
RMSE R2 Model
5.27 0.68 ns
5.34 0.67 P-splines
5.32 0.68 TPRS
```


Simulated data

```
RMSE R2 Model
0.43 0.66 ns
0.15 0.67 P-splines
0.21 0.67 TPRS
```

Two outputs from similar models:

```
summary(model.mgcv)

Formula: y ~ s(x, bs = "cr", k = 7)

Parametric coefficients:
           Estimate   Std. Error t value 
Pr(>|t|)
(Intercept)  -0.09164   0.06232    -1.47   0.142

Approximate significance of smooth terms:
       edf Ref.df F     p-value
s(x) 5.958 5.999 314.1 <2e-16 ***


R-sq.(adj) = 0.654 Deviance explained = 65.6%
GCV = 3.9111 Scale est. = 3.8839 n = 1000
```

```
summary(model)
Call: lm(formula = y ~ ns(x, df = 6), data = df)

Coefficients:
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)      -5.5654 0.3068   -18.140  < 2e-16 ***
ns(x, df = 6)1    9.8984 0.3830    25.846  < 2e-16 ***
ns(x, df = 6)2    2.3910 0.4923     4.857  1.39e-06 ***
ns(x, df = 6)3    7.4688 0.4368    17.097  < 2e-16 ***
ns(x, df = 6)4   -1.7361 0.3808    -4.559   5.79e-06 ***
ns(x, df = 6)5   11.6107 0.7787    14.910  < 2e-16 ***
ns(x, df = 6)6   -4.2501 0.3501    -12.139  < 2e-16 ***

Residual standard error: 1.971 on 993 degrees of freedom
Multiple R-squared: 0.6557, Adjusted R-squared: 0.6536
F-statistic: 315.1 on 6 and 993 DF, p-value: < 2.2e-16
```

Interpretation

- Depending on software output will vary
- Coefficients have no natural meaning/interpretation (eg: odds ratio, risk increase)

```
ns(x, df = 6)1    9.8984 0.3830    25.846  < 2e-16 ***
```

- Standard errors are difficult to interpret
- Testing of hypothesis βj for j function of a base is not meaningful
- Smoothing splines have more complicated forms and penalty make it difficult to obtain a standard error without Bayesian methods
- effective degrees of freedom seem to confuse researchers

How to report results in a clinical paper?

- Splines figure as a main result

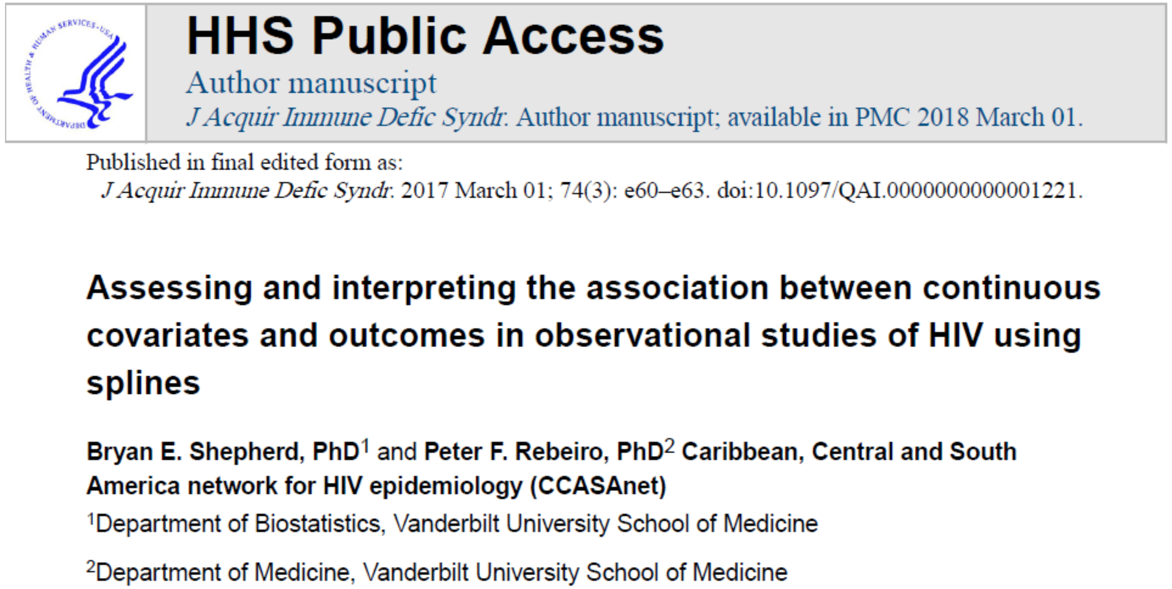
  Often in clinical papers, the statistical reviewer may ask for a spline analysis

  The authors follow the comment but don‘t want to destroy the “nice” clinical conclusion

  So the spline plot is put into the supplement to please the reviewer

- Report typical contrasts

Good example:



Published in final edited form as:

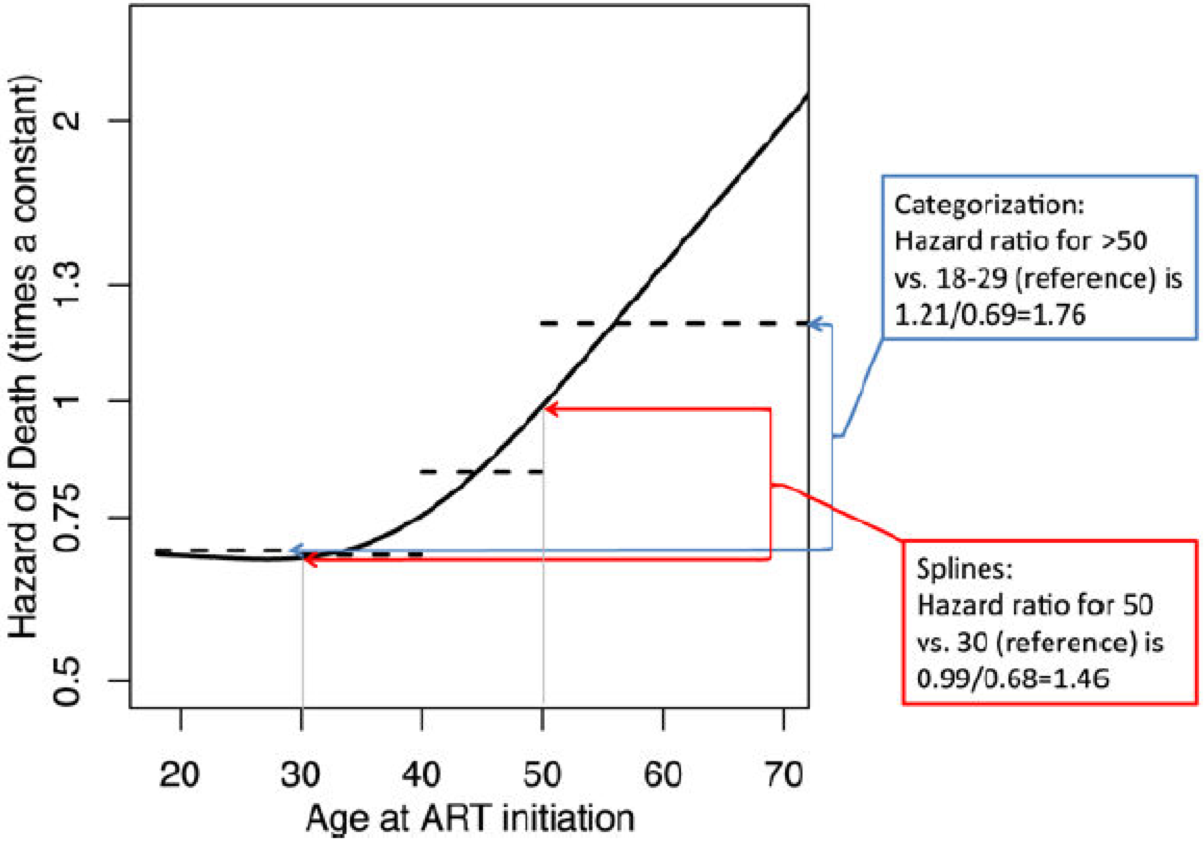
J Acquir Immune Defic Syndr. 2017 March 01; 74(3): e60–e63. doi:10.1097/QAI.0000000000001221.

Assessing and interpreting the association between continuous covariates and outcomes in observational studies of HIV using splines

Bryan E. Shepherd, PhD[1] and **Peter F. Rebeiro, PhD**[2] **Caribbean, Central and South America network for HIV epidemiology (CCASAnet)**

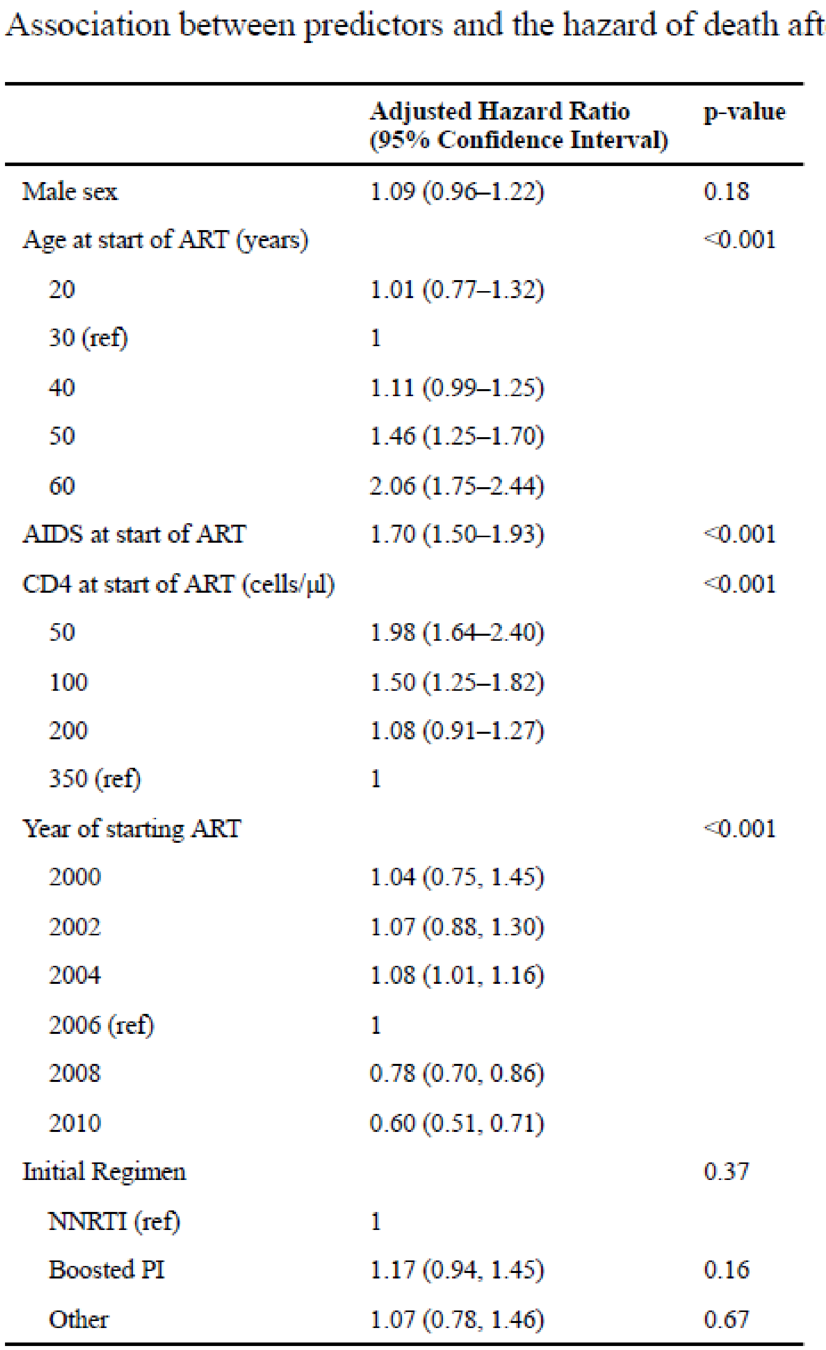
[1]Department of Biostatistics, Vanderbilt University School of Medicine

[2]Department of Medicine, Vanderbilt University School of Medicine



Categorization:
Hazard ratio for >50 vs. 18-29 (reference) is 1.21/0.69=1.76

Splines:
Hazard ratio for 50 vs. 30 (reference) is 0.99/0.68=1.46



Association between predictors and the hazard of death after ART initiation.*

	Adjusted Hazard Ratio (95% Confidence Interval)	p-value
Male sex	1.09 (0.96–1.22)	0.18
Age at start of ART (years)		<0.001
20	1.01 (0.77–1.32)	
30 (ref)	1	
40	1.11 (0.99–1.25)	
50	1.46 (1.25–1.70)	
60	2.06 (1.75–2.44)	
AIDS at start of ART	1.70 (1.50–1.93)	<0.001
CD4 at start of ART (cells/μl)		<0.001
50	1.98 (1.64–2.40)	
100	1.50 (1.25–1.82)	
200	1.08 (0.91–1.27)	
350 (ref)	1	
Year of starting ART		<0.001
2000	1.04 (0.75, 1.45)	
2002	1.07 (0.88, 1.30)	
2004	1.08 (1.01, 1.16)	
2006 (ref)	1	
2008	0.78 (0.70, 0.86)	
2010	0.60 (0.51, 0.71)	
Initial Regimen		0.37
NNRTI (ref)	1	
Boosted PI	1.17 (0.94, 1.45)	0.16
Other	1.07 (0.78, 1.46)	0.67

> Test for non-linearity by contrasting the model fit using splines with a model fit assuming linearity for a specific variable using a likelihood ratio test.

(lack of evidence of non-linearity is not necessarily a reason to simply fit a model assuming a linear relationship)

> With splines, hazard ratios comparing specific contrast can be constructed.

> For example, choose 30 years as the reference age and compute hazard ratios by comparing the hazard of death at select ages with the hazard at 30.

> The hazard ratio for 50 versus 30 years is 0.99/0.68=1.46.

> Any age may be compared to any other age without model re-fitting.

> p-values from likelihood ratio tests with the same number of degrees of freedom as the splines.

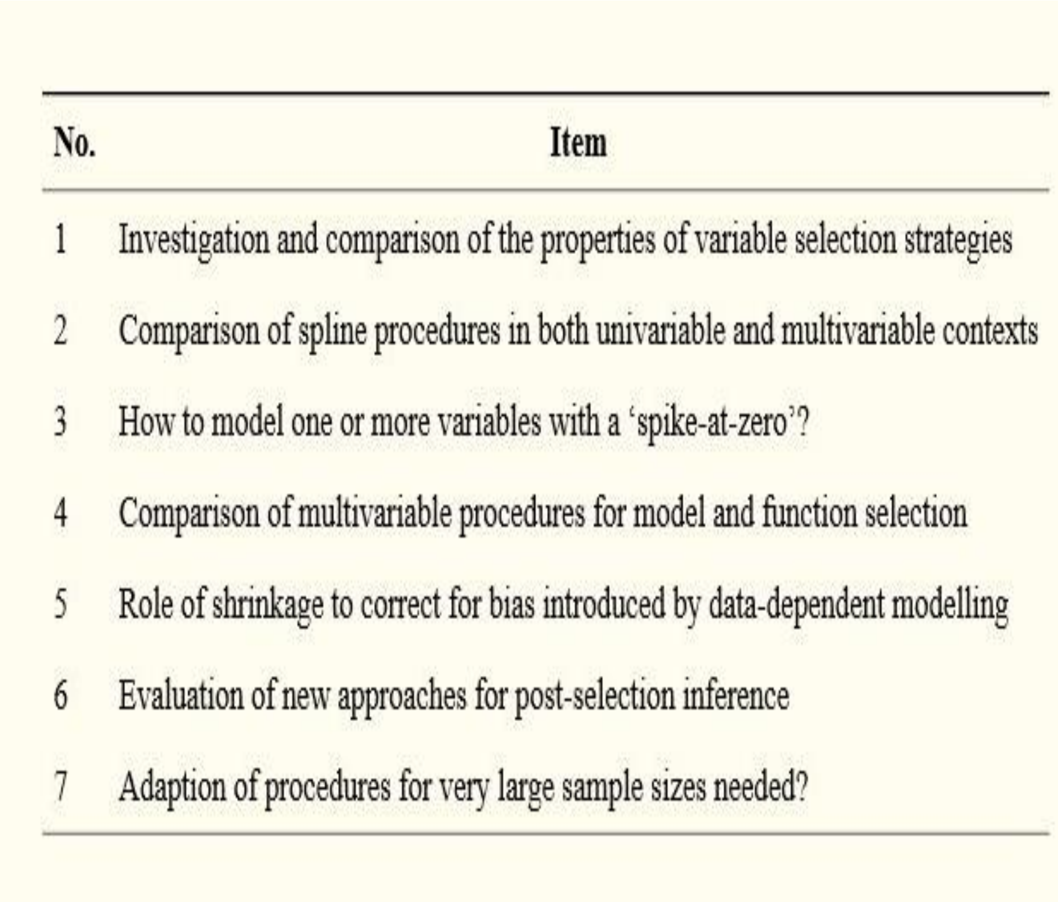
> Correspond to a test that the variable contains predictive information.

On these issues:

Mathematical theory is unlikely to help

Simulation studies are key (Binder et al, StatMed 2013)
However, simulation studies are biased towards the proposed method (Boulesteix et al, BiomJ 2018)
or poorly designed, conducted and reported
(Morris et al, StatMed 2019)

Simulation panel of STRATOS may provide guidance
Experience from comparative analyses with real data sets
Translation to level-1 is needed!



No.	Item
1	Investigation and comparison of the properties of variable selection strategies
2	Comparison of spline procedures in both univariable and multivariable contexts
3	How to model one or more variables with a 'spike-at-zero'?
4	Comparison of multivariable procedures for model and function selection
5	Role of shrinkage to correct for bias introduced by data-dependent modelling
6	Evaluation of new approaches for post-selection inference
7	Adaption of procedures for very large sample sizes needed?

Thanks to all TG2 members!

- Georg Heinze (Austria)
- Willi Sauerbrei (Germany)
- Michal Abrahamowicz (Canada)
- Heiko Becher (Germany)
- Harald Binder (Germany)
- **Daniela Dunkler (Austria)**
- Rolf Groenwold (Netherlands)
- Frank Harrell (U.S.A)
- Nadja Klein (Germany)
- Geraldine Rauch (Germany)
- Patrick Royston (U.K.)
- **Matthias Schmid (Germany)**

And the early career adjunct members

- Michael Kammer (Vienna, Austria)
- Edwin Kipruto (Freiburg, Germany)
- Christine Wallisch (Vienna, Austria)